Agricultural Memoranda No. 319

L. P. Smith

Three helpful statistical methods

1. Chi-squared test for a 2 x 2 contingency table

Having

a	b	a+b
c	d	c+d
a+c	b+d	N

a+b+c+d = N

If $X = \frac{a}{a+b} - \frac{c}{c+d}$

$Y = \frac{a}{a+c} - \frac{b}{b+d}$

Then chi-squared (one degree of freedom)

is XYN /Ref. Nature Dec 22.1951 p. 1087/

P	chi-squared
0.01	6.635
0.02	5.412
0.05	3.841
0.10	2.706
0.20	1.642
0.30	1.074
0.50	0.455

2. Marvin's Method for regression equation when independent variable has equal spacing

Given N values of "y", for equally spaced "x", namely $y_1\ y_2 \ldots\ldots y_N$.

Arrange in 2 equal columns, proceeding down the right hand side and up the left hand side with increasing x. If N is odd, leave the median value by itself at the bottom.
Subtract the right hand column from the left line by line, and multiply the differences by weighting factors N-1, N-3, N-5 etc.
If N is even the last weighting factor is 1; if N is odd the last weighting factor is zero (for the unpaired median value). Add the products to obtain a total T. In symbols

$$(y_N - y_1)(N-1) + (y_{N-1} - y_2)(N-3) \ldots\ldots + (y_{\frac{N}{2}+1} - y_{\frac{N}{2}})(1)$$

or $(y_{\frac{N}{2}+2} - y_{\frac{N}{2}-1})(0)$

Then if M is the mean of the dependent variable and $S = \frac{(N-1)(N)(N+1)}{6}$

the regression is

$$y = M - \frac{(N-1)T}{2S} + \frac{T}{S}x$$

/Values of S

Values of S

N	S	N	S
2	1	12	286
3	4	13	364
4	10	14	455
5	20	15	560
6	35	16	680
7	56	17	816
8	84	18	969
9	120	19	1140
10	165	20	1330
11	220	21	1540

3. Correlation coefficient

Given variables $X_1 \ldots X_r \ldots X_n$ and $Y_1 \ldots X_r \ldots Y_n$
find the mean of X and the mean of Y, then list in two columns the deviations of X_r from mean X and Y_r from mean Y.
namely x_r and y_r

Find A = sum of squares of x_r
B = sum of squares of y_r
C = sum of products or $X_r y_r$
D = square root of product AB

then the coefficient is $\frac{C}{D}$.

Example

The yields of wheat for England and Wales over the years 1950-59 were 21.1, 21.7, 22.6, 24.0, 22.6, 26.7, 24.8, 25.4, 24.6, and 28.8 cwts/acre.

Arrange in the two columns, subtract and multiply by weighting factors

28.8 - 21.1	= 7.7 x 9	=	69.3
24.6 - 21.7	= 2.9 x 7	=	20.3
25.4 - 22.6	= 2.8 x 5	=	14.0
24.8 - 24.0	= 0.8 x 3	=	2.4
26.7 - 22.6	= 4.1 x 1	=	4.1
			110.1 = T

For N = 10. S = 165
Therefore the slope of the regression is $\frac{110.1}{165} = 0.67$
which is the first approximation to the annual yield improvement. Subtracting this annual improvement from the yields and comparing the resulting figures with the potential transpiration over the wheat areas in July and August we find:-

/Adjusted Yield

	Adjusted Yield		July and August P.T	
1950	21.1	Below Average	6.70	Above average
1951	21.0	Below	6.70	Above
1952	21.3	Above	6.70	Above
1953	22.0	Above	7.05	Above
1954	19.9	Below	5.70	Below
1955	23.4	Above	7.60	Above
1956	20.8	Below	6.15	Below
1957	20.7	Below	5.90	Below
1958	19.2	Below	6.40	Below
1959	22.8	Above	7.80	Above
Mean	21.2		6.65	

Arranging this in a 2 x 2 table we find

		Yields Above	Yields Below	
P.T.	Above	4	2	6
	Below	0	4	4
		4	6	10

Chi-squared $= 10 \times (\frac{4}{6} - \frac{0}{4})(\frac{4}{4} - \frac{2}{6})$

$= 10 \times 0.67 \times 0.67 = 4.4$

which is significant at the 0.04 level.

The correlation coefficient can now be found

x	y	x^2	y^2	xy
-0.1	+0.05	0.01	0.0025	-0.005
-0.2	+0.05	0.04	0.0025	-0.010
+0.1	+0.05	0.01	0.0025	+0.005
+0.8	+0.40	0.64	0.1600	+0.320
-1.3	-0.95	1.69	0.9025	+1.235
+2.2	+1.05	4.84	1.1025	+2.310
-0.4	-0.50	0.16	0.2500	+0.200
-0.5	-0.75	0.25	0.5625	+0.375
-2.0	-1.25	4.00	1.5625	+2.500
+1.6	+1.15	2.56	1.3225	+1.840
		14.20	5.8700	8.77
	Square roots	3.77	2.44	

Correlation coefficient $= \frac{8.77}{3.77 \times 2.44} = 0.95$

Meteorological Office
Bracknell
April 1970

G. 04357/VT/4/70/100